DOI: 10.7763/IPEDR. 2012. V49. 12

Research on Online Game Traffic Classification Based on Machine Learning

Zhang Qi and Xiong Wei

College of Medical Information Engineering, Guangdong Pharmaceutical University, Guangzhou 510006, China

jackey0611@hotmail.com, x_w_ei@163.com

Abstract-This paper summarizes online game flow attributes by observing a great number of game data packets and computes their flow feature using Python programming language. Furthermore, we investigate several machine learning algorithms to classify five different online games automatically and correctly, that provide the average accuracy is over 80%. The test results show that machine learning has the better performance than the tradition method in classifying online game traffic.

Keywords-game traffic, machine learning, flow atrributes

1. Introduction

With the advance of network technologies and increasing popularity of online games, online games are gaining more and more attention. Today browser-based games over HTTP are the most popular example of interactive, real-time Internet applications. S.Zander illustrated the identification of game traffic in the Internet is very useful for a number of tasks [1]. Firstly, it is important to find out how much game traffic is in the Internet and how much traffic certain games contribute. Secondly, to provide better QoS for game traffic in the network it is necessary to identify game traffic. In particular, browser game over HTTP shows similar behavior with web surfing traffic that become difficult to identify them correctly. For example, some public computers are not allowed to play online game in library or other public area, but someone still can play it because browser game over HTTP can evade firewall detection easily. Therefore, it is important that online game over HTTP should be separated from another web-based application, such as web or email.

Our work first acquires the characteristics of browser-based games in order to obtain a sufficiently accurate model for web-based game traffic. Then we apply a wide range of better performed machine learning algorithms to separate from different types of online games each other. After evaluation, payload and inter-packet arrival time have a crucial impact on traffic classification. And some of machine learning algorithms are able to identify different online games from each other and have a very high accuracy, e.g. BayesNet algorithm.

2. Related work

In the past, popular methods used to classify network application, such as port number and payload-based identification. However, these techniques have a number of weaknesses. For port-based classification, it relies on mapping application to well-know port numbers. Hence, some applications began using dynamic port numbers and started disguising themselves. For payload-based identification, it can be divided into protocol decoding and signature-based identification. Packets payloads are analyzed to determine whether they contain characteristic signatures of known application. And studies show that these approaches work very well for current Internet traffic [2]. However, there are also some disadvantages. Because of this method relies on specific application data, making it difficult to detect a wide range of applications. Also, the process of creating rules for signature-based classification must often be done by hand, that is very time-consuming.

The limitation of port-based and payload-based analysis have motivated use of transport layer statistics for traffic classification. These classification techniques rely on the fact that different applications have distinct behavior patterns when communicating on the network [3]. Transport layer statistics such as the total

number of packets sent, the ratio of the bytes sent in each direction, the duration of the connection, and the average size of the packets characterize these behaviors.

Machine learning approaches use these transport layer statistics to identify internet applications. T.T.T. Nguyen review 18 significant works on machine learning to IP traffic classification that cover the dominant period from 2004 to early 2007 [4]. Zander proposed in [5] used several flow attributes, including packet length statistics, inter-arrival time statistics and flow duration etc., to identify different types of internet traffic based on AutoClass algorithm. Researchers use classical machine learning algorithms to identify TCP applications [6].

3. online game traffic pattern

According to pattern classification theory, different types of data flow show different behavior characteristic and have unique flow features. Therefore, we summarize online game flow attributes by observing a great number of game data packets in order to build up their own network application data flow model.

3.1 Browser-based Game VS Web Surfing

Both online games over HTTP and web browsing works over TCP connection using port 80. This process is very similar as web browsing. However, playing game behaviors is so different from web browsing. Figure 1 and 2 show the throughputs of game data whose TCP port equals to 80, and web surfing data which generated by Wireshark 1.0.3. Red color curve presents the throughput of HTTP flow. Obviously, the curve trend of game packet is more regular than web browsing. Due to the player must send packets to server regularly based on the game playing rule and server must send the update packets back to player machine too. Unlike web browsing, the web user requests certain web page randomly based on personal surfing Internet habit and preference. Moreover, during the same long period, we can see that the number of packets from web browsing change toughly, the number range from 1400 to 1. While the maximum number of game packets is 500.

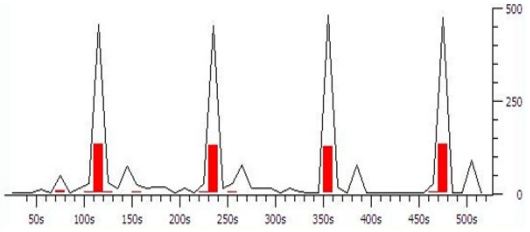

Figure 1 IO graph of online game

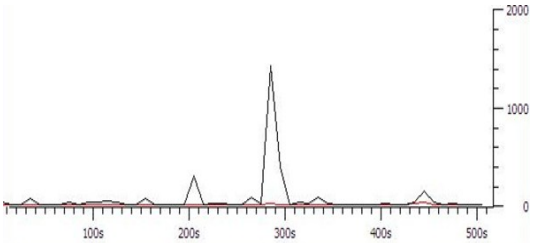

Figure 2 IO graph of web surfing

3.2 Game Flow Attributes

In order to classify different game traffic, it must need to compute flow statistics attributes. We define the flow attributes which derived by [7] in this experiment. In general, the attributes can be summarized as three classes. It includes key attributes, discrete distribution attributes, interactivity attributes.

Key attributes are derived from game flows based on the same port number. And every flow will be identified using 5-tule key, including IP protocol, original IP address, responder IP address, and two corresponding port numbers.

Payload size and inter-packet delay are very important factors for traffic classification. They are belonging to the discrete distribution attributes. Usually, they are represented as bin delimiters. In this study, there are 23 bin delimiters for payload size and 9 bin delimiters for inter-packet delay.

Y. Zhang exploited packet length characteristics, key-stroke packets and command-line packets [8]. Since these two packet length characteristics have great impact on detecting application interactive. Keystroke packets are non-empty small packets, their size is about 60 bytes or less. Especially for online game applications, their feature is sending many short and long packets. In addition, we must know every keystroke payload size, inter-arrival time, and how often the consecutive keystroke packets inter-arrival time fell in the inter-arrival bin. In this work, the keystroke inter-arrival delay time range is 25ms < T < 3000ms. For command-line packets, they have much larger size than keystroke packets carrying 200 bytes or less. And the command-line inter-arrival delay time range is 250ms < T <3000ms. Besides, we quote the definition of conversation and sustained conversation episode from [9].

Especially for game traffic classification, following key flow attributes are needed to compute for identifying game traffic.

- Payload size of first non-empty packet;
- Minimum, maximum, average and standard deviation value of inter-arrival time;
- Minimum, maximum, average and standard deviation value of payload length;
- Total amount of bytes transferred, and total amount of bytes transferred as payload;
- Total number of packets and non-empty packets.

4. Meachine learning algorithms

Machine learning plays an important role in flow classification in today research area. There are different types of machine learning into four basic types, they are classification, clustering, association and numeric prediction [10]. In this work, four supervised classification algorithms are selected. They are C4.5 Decision Tree, Naïve Bayesian Classification, K-Nearest Neighbor, Bayesian Networks.

5. Experimental approach

5.1 Game Selection

To ensure the game data is the most realistic and diverse, all the game data from public game servers that have a large number of real players. There are five popular online games selected, including Yahoo Mahjong, Globulos, QQ Game, Club Marian, and FashionDash.

5.2 Experiment Environment

The machine used to capture online games traffic was an Intel Core 2 machine with 1G RAM, running the Windows XP operating system. Traces of raw traffic were captured by using network sniffer tool, Wireshark 1.0.3.

It is noted that all the games (except game FashsionDash) are online multi-player interactive games, and they are all belong to client-server games. Hence data transmission between the machine and game server is performed using TCP/UDP packets. Table 1 lists the online games measured in the experiment with game protocol and percentage of volume.

Table 1 percentage of volume for each game

Game Class	Protocol	Volume
Mahjong	TCP,HTTP	46.4%
Globulos	TCP,HTTP	9%
QQ game	TCP,UDP,HTTP	13%
Club Marian	TCP,HTTP	11.4%
FashsionDash	TCP,HTTP	20.2%

5.3 Pre-processing data

First of all, capture bidirectional TCP packets using Wireshark. And then, extract the packets information from Wireshark. That is, convert all *.pcap files to *.csv file format. In Wireshark, every row is represented

one packet and each packet is summarized by two attributes, srcPort and dstPort. Then the packets are required to group by flows. One flow is one csv file.

5.4 Statistics Computation

Python is a dynamic object-oriented programming language that can be used for many kinds of software development. Python programming is used here to extract the basic flow attributes from raw data races.

5.5 Machine Learning Software

TANAGRA, ORANGE, and WEKA are all free data mining software. WEKA is conducted here to classify the game traffic datasets. In this experiment, use WEKA to perform data classification using the four algorithms. The experiment has two tests on training datasets. One is testing the classifier by using the whole dataset; another is using dataset to build a prediction model.

6. Experimental Results and Analysis

6.1 Test One Result

Based on upon steps, the whole instances are tested to evaluate how well the prediction model is. In WEKA, choose "cross-validation" (fold=10) in the test option. And it generated good models for data sets through classifier algorithms. Figure 3 lists the accuracy of four classifier algorithms. The overall accuracy is very high, over 82%. In particular, BayesNet algorithm has the highest accuracy (87.12%), while Naïve Bayes Tree has the lowest accuracy percentage (82.58%).

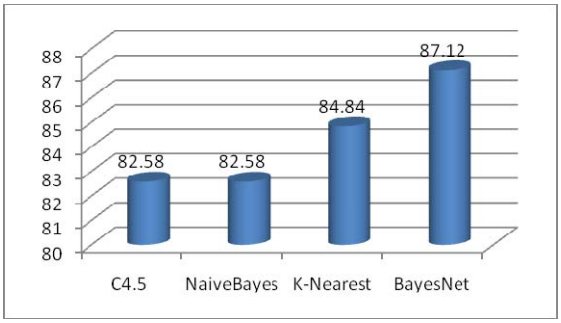

Figure 3 Accuracy per algorithm by using whole data set

6.2 Precision and Recall

To differentiate among classification techniques, there is an important criterion is predictive accuracy. The common method to characterize a classifier's accuracy is through metrics, known as False Positives (FP), False Negatives (FN), True Positives (TP) and True Negatives (TN). Define them as follow. Definition of recall and precision as follows:

$$\text{Recall} = \frac{TP}{(TP+FN)} * 100\%$$

$$\text{Precision} = \frac{TP}{(TP+FP)} * 100\%$$

Figure 4-7 shows the precision and recall using C4.5 algorithm, Naïve Bayesian Classification, K-Nearest Neighbor, and Bayesian Networks, respectively (right bar represents precision, left bar is recall). Figure 4 shows FashsionDash has high percentage of precision and recall by using C4.5 algorithm. In the figure 5, Mahjong, FashsionDash, and Globulos are all have good result of precision using Naïve Bayes algorithm, especially for game Globulos, its precision value is 1. However, these two algorithms can not classify the QQ game instances. On the other hand, by using K-nearest Neighbor algorithm, although no game instances are totally classified correctly, 40% of QQ game instances have been classified correctly. Among four algorithms, BayesNet algorithm has the best result of classify all five game traffic.

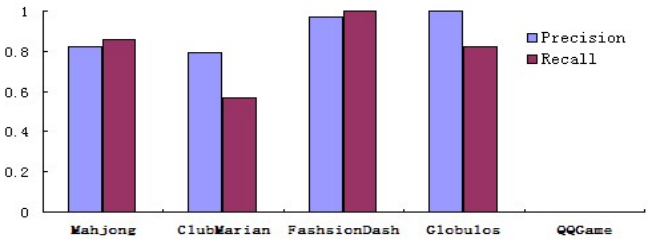

Figure 4 Using C4.5 algorithm

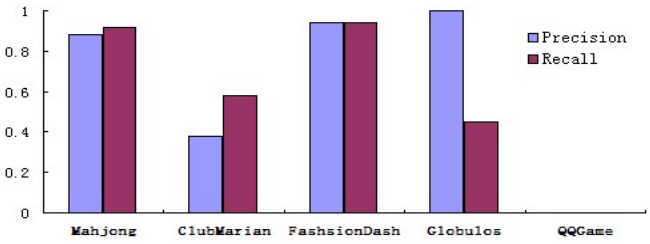

Figure 5 Using Naïve Bayes algorithm

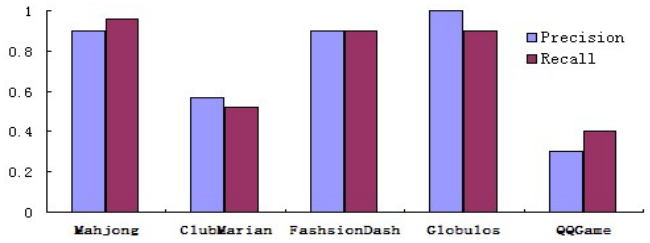

Figure 6 Using K-nearest Neighbor

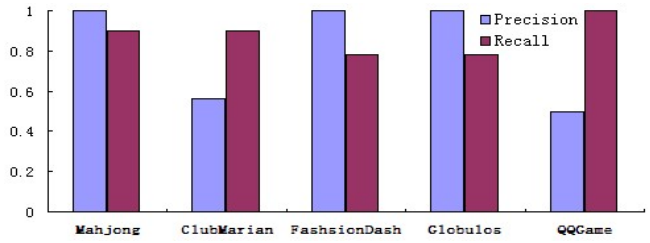

Figure 7 Using BayesNet

6.3Test Two Result

In this test, the whole dataset has divided into training set (75%) and testing set (25%). After testing, it will generate predicted class values automatically. Figure 8 shows the accuracy of four algorithms. The overall accuracy is over 77%. C4.5 decision tree algorithm has the most accurate result (87%), while the least accurate is Naïve Bayesian Algorithm, just occupied 77.42%.

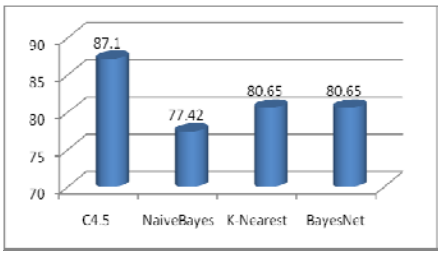

Figure 8 The accuracy of four classifier algorithms

7. Conclusion and Future work

In this work, we study online browser game traffic pattern and compare their traffic with web browsing. Obviously, browser game traffic trend is much more regularly than surfing the internet. Moreover, we use the machine learning techniques to classify five online games. By using classifier learning algorithms, four selected algorithms are all have good predict results that accuracy rate is over 77%. Especially, C4.5 algorithm

has the highest accuracy for predicting the instance, over 87%. While using the model generated by Naïve Bayes algorithm, only 77% instances can be predicted correctly.

In the future, we are going to improve the classification result by trying more suitable algorithms and finding new flow attributes.

8. References

[1] S.Zander, Misclassification of Game Traffic based on Port Numbers: A case study using Enemy Territory, CAIA Technical Report 060410D, April, 2006.

[2] Jeffrey Erman, Martin Arlitt, Anirban Mahanti, Traffic Classification Using Clustering Algorithms, Proc. of ACM SIGCOMM Workshop on Mining Network Data(MineNet), Pisa, Italy, September 2006.

[3] T.Karagiannis, A. Broido, M. faloutsos, and K. claffy. Transport Layer Identification of P2P traffic. In IMC'04, Taormina, Italy, October 25-27, 2004.

[4] T.T.T. Nguyen, G. Armitage, "A Survey of Techniques for Internet Traffic Classification using Machine Learning," IEEE Communications Surveys & Tutorials (to appear 4th edition 2008, accepted Nov 16th 2007)

[5] S. Zander, T. Nguyen, and G. Armitage, "Automated traffic classification and application identification using machine learning," in IEEE 30th Conference on local Computer Networks (LCN 2005), Sydney, Australia, November 2005

[6] M. Crotti, M. Dusi, F. Gringoli, and L. Salgarelli, "Traffic classification through simple statistical fingerprinting," SIGCOMM Comput. Commun.Rev., vol. 37, no. 1, pp. 5–16, 2007

[7] Grenville Armitage & Mark Claypool & Philip Branch, Networking and Online Games, 2006,Chapter 3

[8] Y. Zhang and V. Paxson, "Detecting Backdoors", Proc. Of USENIX Security Symposium, Denver, CO, USA, August 2000.

[9] Ricky A. Bangun & E. Dutkiewicz, Modelling Multi-Player Games Traffic. Information Technology: Coding and Computing, 2000.Proc. International Conference, Las Vegas, NV, USA, March 2000, 228-233.

[10] I, Witten and E. Frank, Data Mining: Practical Machine Learning Tools and techniques with Implementations (second edition), Morgan Kaufmann Publishers, 2005.